dr hab. Maciej Konopiński

Instytut Ochrony Przyrody
Polskiej Akademii Nauk

al. Adama Mickiewicza 33
31-120 Kraków

Kraków, 17 lutego 2025 r.

Recenzja rozprawy doktorskiej pani mgr. Katarzyny Burdy

p.t. „Uwarunkowania obciążenia genetycznego w dzikich populacjach gupików *Poecilia reticulata*"
[ang. „Determinants of genetic load in natural populations of guppy (*Poecilia reticulata*)"]

pod kierunkiem promotora prof. dr hab. Jacka Radwana
oraz promotora pomocniczego dr hab. Mateusza Konczala, prof. UAM

Oceniana rozprawa doktorska została wykonana na Wydziale Biologii Uniwersytetu im. Adama Mickiewicza w Poznaniu, w Pracowni Biologii Ewolucyjnej. Rozprawa została napisana w języku angielskim z wyjątkiem polskiej wersji strony tytułowej i polskiego streszczenia. W dwóch pracach wchodzących w skład rozprawy doktorantka przedstawiła wyniki badań prowadzonych z użyciem metod masowego sekwencjonowania nad czynnikami ewolucyjnymi kształtującymi zmienność w populacjach. W pierwszej pracy, która ukazała się w 2023 roku w czasopiśmie *Molecular Ecology Resources* doktorantka oraz jej promotor pomocniczy, profesor Mateusz Konczal, przedstawili porównanie dwóch metod wykrywania mutacji *de novo*, a niejako ubocznym efektem tego porównania było oszacowanie tempa mutacji u gupików, hodowanych na UAM. Jest to praca o charakterze metodycznym mająca tylko pośredni związek z tematem rozprawy doktorskiej. Tempo mutacji obliczone w tej pracy zostało użyte w drugiej pracy wchodzącej w skład rozprawy. Zgodnie z przedstawionym oświadczeniem o autorstwie pani Katarzyna Burda współtworzyła koncepcję badań, przeprowadziła większość analiz laboratoryjnych, bioinformatycznych oraz statystycznych, oraz współtworzyła manuskrypt artykułu a następnie brała udział rewizji artykułu, natomiast rola promotora pomocniczego ograniczyła się do etapu planowania badań, jednego z etapów obróbki bioinformatycznej oraz współtworzenia manuskryptu artykułu i jego rewizji. Druga praca, bezpośrednio dotycząca tytułu rozprawy, ma postać wieloautorskiego manuskryptu w formacie artykułu naukowego, w którym opisano wyniki badań genomowych nad gupikami występującymi w ich środowisku naturalnym na wyspach archipelagu karaibskiego Trinidad i Tobago. Pani Katarzyna wzięła udział we wszystkich etapach powstawania tej pracy z wyjątkiem jednego z etapów obróbki bioinformatycznej. Zarówno udział we wszystkich istotnych etapach tworzenia obu prac, oraz pierwsza pozycja na listach autorów świadczą o wiodącym wkładzie doktorantki w powstanie artykułu i manuskryptu składających się na ocenianą rozprawę doktorską.

Pierwsza z tych prac ma charakter techniczny. Wykrywanie mutacji *de novo* tak na prawdę stało się możliwe dopiero dzięki rozwojowi technik masowego sekwencjonowania nowej generacji (ang. *Next Generation Sequencing*, NGS). Substytucje w genomach eukariotycznych zdarzają się dosyć rzadko – rzędu raz na setki milionów lub nawet miliardy nukleotydów na pokolenie. Oznacza to, że żeby wykrywać pojedyncze mutację u organizmów rozmnażających się płciowo, należy zsekwencjonować DNA o długości miliardów par zasad. Niestety wyniki sekwencjonowania metodami NGS są obarczone stosunkowo sporą liczbą błędów. Przyjmuje się, że akceptowalny poziom błędu wynosi między 10^{-3} a 10^{-4}, czyli poszukując pojedynczej mutacji mamy do czynienia z milionami błędnych odczytów. Błędy te są odsiewane na etapie obróbki bioinformatycznej. Istnieje jednak dylemat wyboru kryteriów filtrowania – zbyt restrykcyjne mogą doprowadzić do „odsiania" prawdziwych mutacji (wynik fałszywie negatywny), podczas gdy zbyt łagodne mogą prowadzić do uznania błędu sekwencjonowania za nową mutację (wynik fałszywie pozytywny).

W artykule stanowiącym pierwszy rozdział rozprawy doktorantka razem ze swoim promotorem dokonali porównania dwóch metod wykrywania mutacji: „twardego filtrowania" (ang. *„hard filtering*) oraz filtrowania z użyciem metody uczenia maszynowego (ang. *machine learning*, ML). [Nawiasem mówiąc, skrót ML jest trochę mylący w kontekście genetyki populacyjnej, ponieważ jest on powszechnie kojarzony raczej z metodami *Maximum-Likelyhood* niż z uczeniem maszynowym. To powinno być wyjaśnione już w streszczeniu, gdzie skrót pojawia się po raz pierwszy.] Do porównania skuteczności obu metod wykorzystano dane z sekwencjonowania dwóch rodzin gupików hodowanych w Pracowni Biologii Ewolucyjnej UAM. Zsekwencjonowano łącznie genomy 24 osobników, w tym genomy obu par rodziców oraz wybranych 10 osobników potomnych z każdej pary. Znaleziono i pozytywnie zweryfikowano łącznie 24 mutacje w 22 miejscach, przy czym metoda wykorzystująca uczenie maszynowe pozwoliła na wykrycie trzech dodatkowych mutacji, które nie zostały wykryte przy zastosowaniu *hard filtering*. Wyniki oszacowania ogólnego tempa mutacji nie różniły się istotnie między obiema metodami.

Praca jest napisana zrozumiałym językiem, posiada jasno przedstawiony cel, a wyniki zostały pokazane i przedyskutowane w sposób klarowny. Pani Katarzyna Burda wykazała, że metody uczenia maszynowego stanowią bardzo dobrą alternatywę dla wcześniejszych historycznie metod „twardego filtrowania". Wyższa wykrywalność mutacji *de novo* pozwala na bardziej precyzyjne wyliczenie tempa mutacji, co ma znaczenie w badaniach nad mutacjami w kontekście ewolucji genomu.

Chociaż artykuł był recenzowany przed publikacją i opublikowany w bardzo dobrym czasopiśmie znalazłem w nim kilka drobnych uchybień i nieścisłości, które wymienię poniżej. Brakuje mi w pracy choćby wzmianki o innych rodzajach mutacji. Słowo mutacja jest tu używane jako synonim substytucji, a przecież oprócz substytucji w genomach zdarzają się też inwersje, oraz mniejsze lub większe insercje i delecje. Wspomnienie o tym we wstępie nie zajęłoby wiele miejsca, a umieszczałoby przedstawione wyniki w szerszym kontekście. Brakuje mi również pokazania w wynikach i dyskusji informacji na temat lokalizacji znalezionych mutacji – czy to są geny, introny czy inne fragmenty genomu, a jeśli któraś z mutacji występuje w genie, jaki skutek wywołuje. Jedyne co wiadomo to, że pominięto mutacje w elementach powtarzalnych, co wynika z przyjętych założeń. Genom gupików został poznany i adnotowany dosyć dawno, dlatego nie wymagałoby to wiele dodatkowego wysiłku, zwłaszcza, że w drugiej pracy, która wchodzi w skład rozprawy, taka informacja się znalazła. Kolejna uwaga dotyczy źródła materiału genetycznego - pozyskano go z ogonów, czyli *de facto* głównie z komórek nabłonka, które nie stanowią linii zarodkowej (ang. *germline*). Być może lepszym wyjściem byłoby sekwencjonowanie całych osobników tuż po porodzie, aby wykluczyć mutacje somatyczne. Uchybieniem technicznym można nazwać brak zastosowania kroku *base recalibration* zalecanego przy poszukiwaniu polimorficznych nukleotydów z użyciem pakietu *GATK*. Ten krok, choć mocno czasochłonny u organizmów niemodelowych, pozwala na wyłowienie większej liczby polimorfizmów (a tym samym mutacji *de novo*) oraz skorygowanie części błędów. Zdaję sobie sprawę, że może za wyjątkiem ostatniej uwagi dwie wcześniejsze nie mają żadnego wpływu na to co jest głównym tematem pracy, ale pisząc artykuł warto pamiętać, że informację, która wcale nie była głównym celem publikacji, czytelnicy mogą uznać za istotną. Dla wielu badaczy jednym z najciekawszych elementów tej pracy będzie nowe oszacowane tempa mutacji u gupików. Z resztą temu aspektowi poświęcono całkiem sporo miejsca w dyskusji. Moim zdaniem należało w pracy wspomnieć o kilku ograniczeniach przedstawionego oszacowania. Na przykład warto byłoby napisać o efekcie *time-dependecy*, aby uczulić na ograniczenia stosowania opublikowanego tempa mutacji w badaniach obejmujących dłuższe okresy (tj. dziesiątki lub setki tysięcy lat). Część z nowo powstałych mutacji zmniejsza dostosowanie osobników i jest dosyć szybko usuwana przez dobór. Dlatego oszacowania mutacji wprost nie mogą być stosowane dla analiz dotyczących dłuższych okresów niż kilka tysięcy lat. Brakuje mi również kilku podstawowych informacji na temat samej hodowli – np. od ilu pokoleń, lub choćby od ilu lat populacja znajduje się w laboratorium oraz jak liczna jest ta populacja. Jak zauważają sami autorzy tempo mutacji również może podlegać doborowi naturalnemu i nie można wykluczyć, że hodowla sprzyja przyspieszeniu lub spowolnieniu tempa mutacji (np. przez brak w hodowli patogenów, które mogą wymuszać szybsze różnicowanie genów odpornościowych, albo przez samo utrzymywanie temperatury odbiegającej od tej panującej w warunkach naturalnych).

Drugi rozdział rozprawy to manuskrypt artykułu na temat zmian w obciążeniu genetycznym (ang. *genetic load*) w odniesieniu do demografii gatunku. W manuskrypcie podjęto próbę odpowiedzi na pytanie jakie jest rozmieszczenie obciążenia genetycznego w zależności od poziomu zmienności oraz w procesie inwazji gupików w rzekach Północnego Trynidadu oraz Tobago, czyli w naturalnym zasięgu występowania tego gatunku. Temat obciążenia genetycznego jest niezwykle istotny z punktu widzenia genetyki konserwatorskiej i cieszy mnie, że doktorantka postanowiła się z nim zmierzyć.

We wstępie doktorantka przedstawia główne założenia i opisuje dosyć obszerny stan wiedzy na temat naturalnych populacji gupików. W tej części jest kilka niezręcznych sformułowań. Na przykład: „Apart from conservation context (...) is a biological invasion" – przecież gatunki inwazyjne są jak najbardziej w kontekście konserwatorskim i mają ogromne znaczenie dla globalnego spadku bioróżnorodności. „Initial stages of invasion" nie koniecznie charakteryzują się małą liczbą osobników. Ostatni akapit wstępu jest zbyt szczegółowy i należałoby go raczej przenieść do metod.

W celu przetestowania postawionych hipotez pani Katarzyna Burda zebrała próby materiału genetycznego z 14 miejsc na terenie 8 rzek należących do dwóch izolowanych zlewniach w północnej części Trinidadu oraz z 2 rzek na Tobago. Mapka zawarta w manuskrypcie wymaga dopracowania – na mapce brakuje skali a jeszcze lepiej linijki, brakuje szerszego planu, który obejmowałby również rzeki z Tobago, i choć fragment Ameryki Południowej, czy Morze Karaibskie, brakuje zaznaczonych lokalizacji, z których pochodzą próbki. Niby w tabeli w materiałach uzupełniających podano koordynaty, ale ta informacja jest nieczytelna bez zastosowania programów mapowych. Dodatkowo warto byłoby pokazać na mapie zasięg występowania gupików oraz większe bariery dla ekspansji/migracji, np. duże wodospady, czy tamy, jeśli takie zostały zbudowane na tych rzekach co może pomóc we własnej reinterpretacji wyników przez czytelnika. Poza tym tabela ze spisem lokalizacji i liczebności prób powinna moim zdaniem znaleźć się w treści artykułu, a nie jako *Supplementary table,* ponieważ jest to jedna z ważniejszych informacji technicznych w opisie metod.

Doktorantka analizowała sekwencje genomowe od 190 osobników. Nie wiadomo kto wykonał biblioteki do sekwencjonowania, natomiast prawie całą obróbkę bioinformatyczną wykonała doktorantka. W zasadzie nie jest jasne dlaczego akurat etap wywołania loci polimorficznych z użyciem *bcftools* wykonywał pan profesor Konczal. Nie umniejsza to ogromu pracy jaki doktorantka włożyła w realizację badań, ale nie jest to zadanie trudniejsze od pozostałych etapów obróbki, a bywa ono ważnym etapem analiz bioinformatycznych. Być może to właśnie skutkiem tego podziału zadań rozdział dotyczący metod jest trochę chaotyczny i sprawia wrażenie pośpiesznie złożonego z fragmentów napisanych przez kilku autorów. Analizy bioinformatyczne zostały niepotrzebnie rozrzucone w kilku podrozdziałach, co utrudnia czytanie pracy. Ten etap analiz, na równi z samym sekwencjonowaniem, jest tylko instrumentem do uzyskania właściwych danych, a dopiero analizy statystyczne odróżniają genetyka populacyjnego od biologa molekularnego. Wyodrębnienie analiz statystycznych ułatwiłoby prześledzenie toku rozumowania autorki. Na przykład dało by szansę wyjaśnienia dlaczego w pracy użyto dwóch programów robiących praktycznie to samo: *bcftools* i *HaplotypeCaller*. Podobnie jak w przypadku poprzedniej publikacji przy obróbce danych z użyciem *GATK* doktorantka pominęła krok *base recalibration*, który mógłby poprawić jakość danych i liczbę uzyskanych SNPów.

Nie rozumiem dlaczego doktorantka nie wykorzystała wszystkich możliwości jakie niesie za sobą wyznaczenie w pierwszej pracy tempa mutacji Zostało ono użyte jedynie jako przelicznik przy szacowaniu efektywnej wielkości populacji. Mam zresztą wątpliwości, czy akurat te obliczenia są uprawnione. Prosty, klasyczny wzór na obliczanie efektywnej wielkości populacji z poziomu zmienności oraz tempa mutacji ma zastosowanie jedynie do populacji w równowadze między dryfem a nowymi mutacjami bez migracji między populacjami. To założenie jest bardzo trudne do spełnienia. Populacja z rzeki Turure z pewnością go nie spełnia. Można mieć również wątpliwości co do pozostałych populacji gupików. Pod koniec plejstocenu poziom morza był niższy o ponad 100 m. Wiele obecnie izolowanych populacji było wtedy połączonych, prawdopodobnie istniało lądowe połączenie między Trynidadem a Tobago, więc sytuacja całego gatunku była zupełnie inna niż obecnie. Trudno powiedzieć, czy dzisiejsze populacje zdołały osiągnąć równowagę wymaganą przez to równanie. W dyskusji termin *effective population size* bywa używany

zamiennie z *genetic variation* (np. przy cytowaniu ryciny 4 w dyskusji), które to określenie w tym kontekście wydaje się być znacznie bardziej na miejscu. Mając sekwencje całych genomów osobników można było wykorzystać znacznie bardziej zaawansowane podejście, np. obliczyć wielkość populacji na podstawie nierównowagi sprzężeń (ang. *linkage disequlibrium*). Jeśli zaś chodzi o porównanie wyników z populacjami z Tobago można mieć wątpliwości, czy próba 5-ciu osobników jest wystarczająca do wyciągnięcia jakichkolwiek uprawnionych wniosków. Z jednej strony mamy do czynienia z danymi genomowymi i różnorodność nukleotydowa uśredniona dla 5-ciu genomów może rzeczywiście reprezentować całą populację, jednak przy tak małej próbie każde odstępstwo od losowości (np. dwa spokrewnione osobniki w próbie) może mieć ogromny wpływ na uzyskany wynik. Niestety w pracy brakuje dokładniejszych informacji na temat populacji z Tobago oraz miejsc pobrania prób.

Tempo mutacji należałoby wykorzystać przy analizach filogeograficznych z zastrzeżeniem, o którym pisałem wcześniej, czyli ograniczając konkluzje do wąskich ram czasowych, rzędu kilku-kilkunastu tysięcy lat. To zresztą jest wielki nieobecny tych analiz. Zdaję sobie sprawę, że zarówno tytuł jak i główny temat drugiej pracy dotyczy badań nad obciążeniem genetycznym, jednak zbadanie historycznych związków między populacjami przy pomocy programów takich jak *Treemix*, czy *Admixtools* stanowiłoby doskonałe uzupełnienie wyników z PCA czy *Admixture*.

Przez cały manuskrypt przewija się wątek hipotezy o „czyszczeniu ze szkodliwych wersji" (ang. *genetic purging*), która mówi, że gatunki inwazyjne przechodząc u początku inwazji wąskie gardło liczebnościowe pozbywają się części obciążenia genetycznego, co miałoby powodować zwiększenie dostosowania i przyczynić się do sukcesu inwazji. Jak już zauważono we wstępie manuskryptu, ta hipoteza rzadko znajduje potwierdzenie w badaniach, dlatego nie dziwi, że i doktorantce nie udało się odnaleźć podobnego wzoru. Niestety w obecnych czasach negatywna weryfikacja nie jest zbyt doceniana przez wydawców, dlatego być może przed publikacją warto byłoby przeorientować artykuł tak, aby nie była to centralna oś artykułu. Dodatkowo nie warto się upierać przy używaniu populacji z Turure do testowania hipotezy o gromadzeniu się obciążenia genetycznego na froncie inwazji. Populacja, która według doktorantki miała być tą inwazyjną, została w rzeczywistości wypuszczona w innej części zasięgu tego samego gatunku, dlatego można się spierać, czy jest to rzeczywiście sytuacja odpowiadająca inwazji. Osobniki z miejsca o wysokiej presji drapieżników wypuszczono, w miejscu o niskim drapieżnictwie, w górnym biegu rzeki, gdzie gupiki wcześniej nie występowały. Jeśli nie było tam wcześniej gupików, można się domyślać, że to miejsce jest odizolowane od reszty zasięgu jakimś wodospadem lub inną barierą. Jeśli nowa populacja znalazła tam odpowiednie warunki do rozrodu, i jednocześnie została pozbawione presji drapieżników, mogła stać się rodzajem źródła, które przez kilka dekad „dolewało" genów do puli populacji znajdujących się poniżej. Taki niekończący się strumień osobników mógł w niedługim czasie zdominować pule genetyczne populacji znajdujących w dolnym biegu rzeki, zwłaszcza, jeśli osobniki mieszańcowe miały niższe dostosowanie. Doktorantka powinna spróbować wykorzystać programy takie jak *DIYABC-RF* lub *fastsimcoal2*, które są skuteczniejszym narzędziem do analizy zmian demograficznych niż próba interpretacji kolorów na wykresach z *Admixture*.

Inne uwagi do drugiego rozdziału:

- Opisy rycin i tabel bywają niefortunne. Nie powinny one zaczynać się od np. „Figure presenting..." albo „Results of statistical tests performed."

- Na rysunku *Supplementary figure 2* brakuje liter na oznaczenie poszczególnych wykresów, i wystarczył by jeden podpis osi x na dole.

- W metodach wspomniano, że sekwencjonowanie wykonano w dwóch rzutach (w 2019 i 2022 roku), ale nie ma żadnych dodatkowych informacji na ten temat. Choć metody NGS są znacznie lepiej wystandaryzowane niż np. analiza długości loci mikrosatelitarnych, tego rodzaju sytuacje bywają źródłem błędu systematycznego. Czy osobniki ze środkowego biegu rzeki Turure, które nie pasują do reszty tej populacji, były sekwencjonowane w jednym rzucie z pozostałymi?

- W manuskrypcie zdarzają się literówki (np. w odnośniku do strony Github znajdującym się na początku metod), błędy formatowania (np. brak wyróżnienia nazw funkcji i programów), czy nieistotne informacje jak użycie funkcji *cbind* do połączenia zbiorów danych (dla niewtajemniczonych: to jest ten sam poziom komplikacji jak skopiowanie komórek w Excelu), ale ponieważ podejrzewam, że przejdzie on poważną reorganizację, nie będę zanudzał szczegółami czytelników tej recenzji.

Mimo tych wszystkich krytycznych uwag, obie prace stanowią moim zdaniem ważny wkład w rozwój genomiki ewolucyjnej. Dzięki łatwości utrzymania w hodowli gupiki są jednym z organizmów modelowych i odgrywają dużą rolę w biologii ewolucyjnej np. w badaniach nad doborem płciowym. Każda praca poszerzająca wiedzę na temat mechanizmów kształtujących ich zmienność genetyczną ma dużą wartość naukową. Warto podkreślić, że przeprowadzenie badań i analiz bioinformatycznych i statystycznych na danych całogenomowych nie było z pewnością łatwym zadaniem dla osoby na tym etapie kariery naukowej. Chociaż manuskrypt stanowiący drugi rozdział rozprawy jest miejscami chaotyczny, a potencjał zgromadzonych danych nie został w pełni wykorzystany, to jednak świadczy on o tym, że Pani Katarzyna potrafi krytycznie spojrzeć na własne wyniki i nie próbuje ich dopasowywać do wcześniejszych założeń. Błędy i niezręczności w manuskrypcie wynikają prawdopodobnie z braku doświadczenia na tym etapie kariery naukowej oraz być może z niewystarczającego wsparcia promotorów doktorantki. Jestem pewien, że zdobywając dalsze doświadczenie i poszerzając swoją wiedzę w dziedzinie genomiki, Pani Katarzyna ma potencjał, aby rozwinąć się jako dojrzały naukowiec.

Podsumowując stwierdzam, że rozprawa doktorska pani mgr Katarzyny Burdy spełnia w mojej ocenie warunki określone w art. 187 ust. 1-2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2024 r. poz. 1571). Składam zatem wniosek do Rady Naukowej Dyscypliny Nauki Biologiczne Uniwersytetu im. Adama Mickiewicza w Poznaniu o dopuszczenie pani mgr Katarzyny Burdy do dalszych etapów przewodu doktorskiego.

Maciej Konopiński